Part 4: Regression

Sam Tyner

TBD

Textbook

These slides are based on the book *OpenIntro Statistics* by David Diez, Christopher Barr, and Mine Çetinkaya-Rundel

The book can be downloaded from https://www.openintro.org/stat/textbook.php

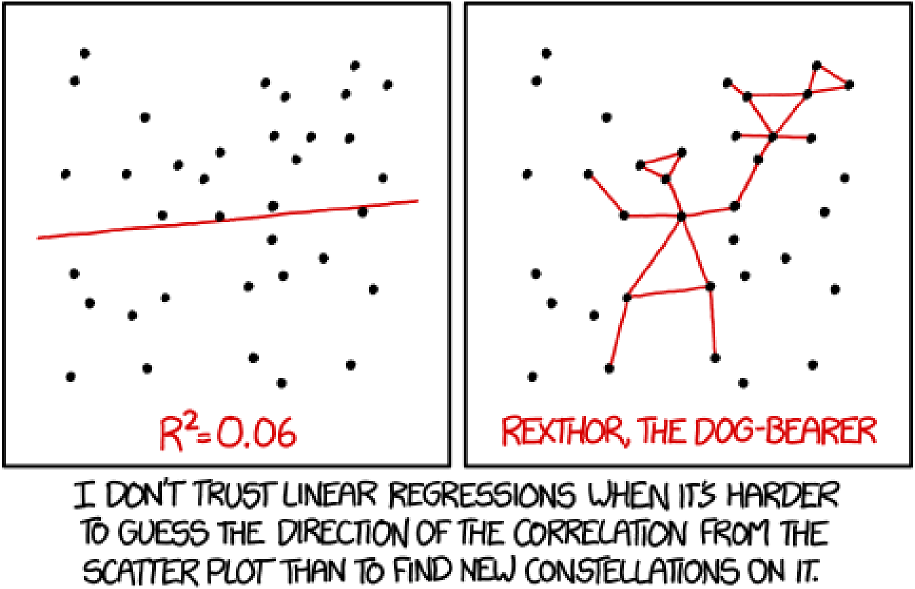
Part 4 Corresponds to Chapters 7, 8 of the text. Sections 4.1-4.3 correspond to chapters 7, 8.1, 8.4 of the text.

Outline

- Simple Linear Regression (4.1)
- Multiple Regression (4.2)
- Logistic Regression (4.3)

Section 4.1: Simple Linear Regression

What is regression?



Formula for a line

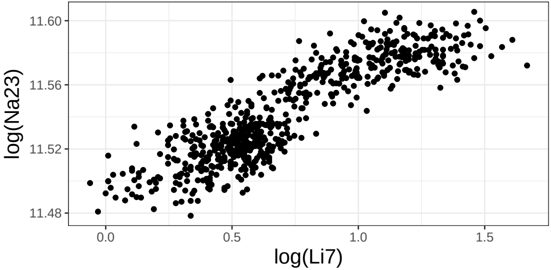
$$Y = a \cdot X + b$$

- Y: dependent variable (response)
- X: independent variable (predictor)
- a: slope (for every 1 unit increase in X, Y increases by a)
- b: intercept (when $X = 0$, $Y = b$)
- *Deterministic*: knowledge of a, X, b means you know Y

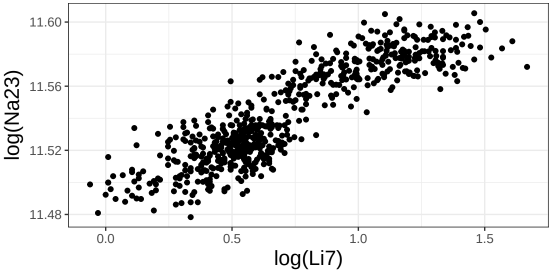
Linear Regression

Why?

- Have two variables, Y, X and we think that the value of Y *depends on* the value of X
- Why would we think that? Maybe we have previous knowledge or we looked at a scatterplot of the data



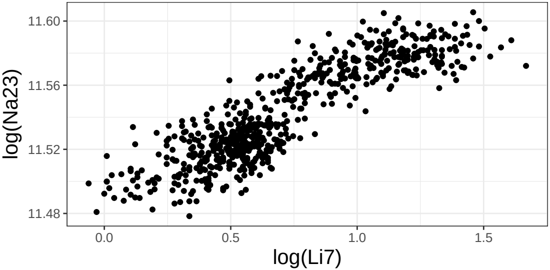
Linear Regression



Guesstimate the slope and intercept of a line through this data

- a: slope = ?
- b: intercept = ?

Linear Regression



Guesstimate the slope and intercept of a line through this data

- a: slope $\approx$ 0.08
- b: intercept $\approx$ 11.48

Best fit line

The *best fit line* is the equation $Y = a \cdot X + b$ with values a, b that minimize the **sum of squared residuals**. What does that mean?

- Write the best fit line as $\hat{Y} = \beta_0 + \beta_1 \cdot X$.

Best fit line

The *best fit line* is the equation $Y = a \cdot X + b$ with values a, b that minimize the **sum of squared residuals**. What does that mean?

- Write the best fit line as $\hat{Y} = \beta_0 + \beta_1 \cdot X$.
- $\hat{Y}$ is the predicted value of Y by the best fit line

Best fit line

The *best fit line* is the equation $Y = a \cdot X + b$ with values a, b that minimize the **sum of squared residuals**. What does that mean?

- Write the best fit line as $\hat{Y} = \beta_0 + \beta_1 \cdot X$.
- $\hat{Y}$ is the predicted value of Y by the best fit line
- **Residual** - what is "left over" after the prediction.

Best fit line

The *best fit line* is the equation $Y = a \cdot X + b$ with values a, b that minimize the **sum of squared residuals**. What does that mean?

- Write the best fit line as $\hat{Y} = \beta_0 + \beta_1 \cdot X$.
- $\hat{Y}$ is the predicted value of Y by the best fit line
- **Residual** - what is "left over" after the prediction.
- Denote residual for observation i by $e_i = Y_i - \hat{Y}_i$

Best fit line

The *best fit line* is the equation $Y = a \cdot X + b$ with values a, b that minimize the **sum of squared residuals**. What does that mean?

- Write the best fit line as $\hat{Y} = \beta_0 + \beta_1 \cdot X$.
- $\hat{Y}$ is the predicted value of Y by the best fit line
- **Residual** - what is "left over" after the prediction.
- Denote residual for observation i by $e_i = Y_i - \hat{Y}_i$
- We are minimizing $\sum_{i=1}^{N} e_i^2$

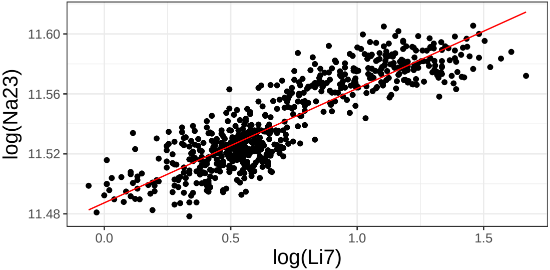
Calculating the best fit line

- $\beta_1 = \frac{s_y}{s_x} \cdot r$
- s_y: standard deviation of the data observations Y
- s_x: standard deviation of the data observations X
- r: correlation between the observations X, Y (measure of association between X, Y)
- $\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}$

Do it in R

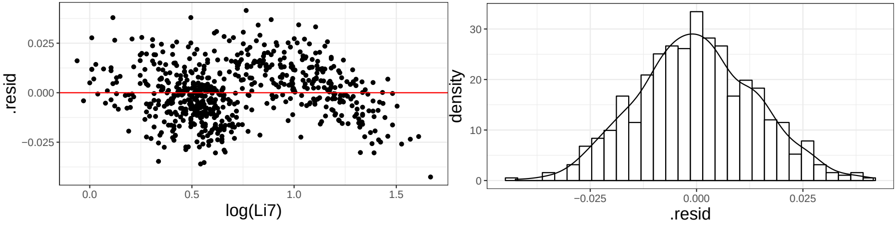
```r
bfl <- lm(data = glass, log(Na23) ~ log(Li7))
coef(bfl)
```

```
## (Intercept)    log(Li7)
## 11.48732735  0.07632503
```



Residual Plot

Look at the residuals e by the values of X:



Want to see a random scatter of points above and below 0

R^2: how well does X explain Y?

R^2, the **coefficient of determination** defines how much of the variability in Y is explainable by the values of X.

$$R^2 = 1 - \frac{Var(e)}{Var(y)}$$

Example:

- $Y = \log(Na23)$. $Var(Y) = 0.00089$
- $e_i = Y_i - \hat{Y}_i = Y_i - (11.487 + 0.0763 \cdot X_i)$. $Var(e) = 0.00019$
- $\frac{Var(e)}{Var(y)} = \frac{0.00019}{0.00089} = 0.2138$
- $R^2 = 1 - \frac{Var(e)}{Var(y)} = 1 - 0.2138 = 0.7862$

78.62% of the variability in Y is explained by the value of X.

Section 4.2: Multiple Regression

Multiple Regression = Multiple Predictors

Multiple regression is the same general idea as simple linear regression, but instead of one predictor variable, X, we have two or more: $X_1, X_2, \ldots, X_p$ where p is the number of predictor variables.

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p$$

β_0 (intercept) and $\beta_1, \ldots, \beta_p$ are called *coefficients*

- β_0 is the value of $\hat{Y}$ when ALL Xs are 0
- $\beta_k, k \in \{1, 2, \ldots, p\}$ is the amount that Y increases when X_k increases by 1 unit, and *all other X values are held constant*

Why?

Why Multiple Regression?

- May know that more than 1 variable affects the value of Y (background knowledge)
- More predictors generally means better fit, better predictions

Example: Add log value of Neodymium (common glass additive) to the model

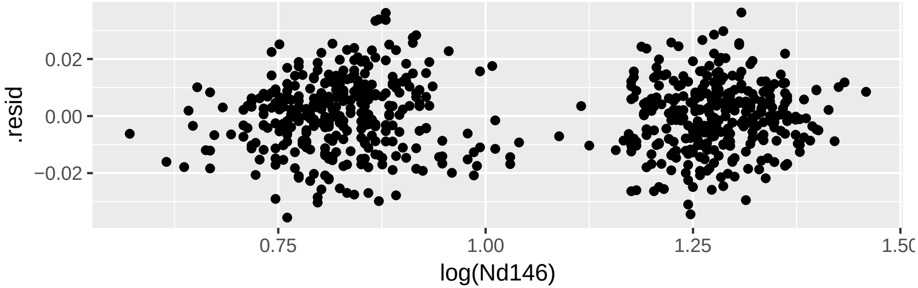
```
blfm2 <- lm(log(Na23) ~ log(Li7) + log(Nd146), data = glass)
blfm2
```

```
##
## Call:
## lm(formula = log(Na23) ~ log(Li7) + log(Nd146), data = glas
##
## Coefficients:
## (Intercept)     log(Li7)   log(Nd146)
##    11.53947      0.05774     -0.03699
```

Multiple Regression Example

Example: $\log(Na23) = \beta_0 + \beta_1 \cdot \log(Li7) + \beta_2 \cdot \log(Nd146)$

Residuals:



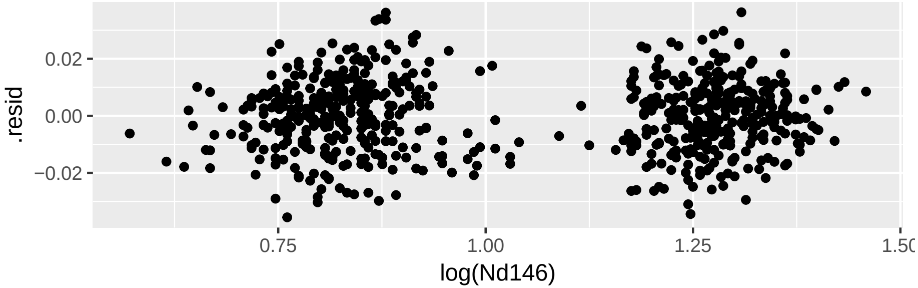
$R^2 = 1 - \frac{Var(e)}{Var(y)} = 1 - \frac{0.000155}{0.00089} = 0.8258$

What do you think of this model?

Multiple Regression Example

Example: $\log(Na23) = \beta_0 + \beta_1 \cdot \log(Li7) + \beta_2 \cdot \log(Nd146)$

Residuals:



$$R^2 = 1 - \frac{Var(e)}{Var(y)} = 1 - \frac{0.000155}{0.00089} = 0.8258$$

Better fit than the simple linear regression, but we uncovered a new pattern: two distinct groups of residuals

Another multiple regression

There are 2 manufacturers in the glass data, so we'll add the manufacturer as a variable in the model:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$$

- $Y = \log(Na23)$, $X_1 = \log(Li7)$, $X_2 = \log(Nd146)$, $X_3 =$ manufacturer.

```
blfm3 <- lm(log(Na23) ~ log(Li7) + log(Nd146) + mfr , data = glass)
blfm3
```

```
##
## Call:
## lm(formula = log(Na23) ~ log(Li7) + log(Nd146) + mfr, data = glas
##
## Coefficients:
## (Intercept)     log(Li7)   log(Nd146)        mfrM2
##   11.523955     0.057276    -0.024866     0.006241
```

Section 4.3: Logistic Regression

Different types of responses

In linear & multiple regression, we have a *continuous* numerical response variable. However, this is not always the case.

Often, we want to determine whether the response belongs in one of two categories.

Examples:

- Is an email spam or not?
- Is a glass fragment from manufacturer 1 or 2?
- Will a juror say the defendant in a case is guilty or not guilty?

Logistic regression

Idea:

When the outcome is one of 2 options, you select the “success” outcome and model the response as binary.

- If $Y_i = 1$ the response is a “success”, if $Y_i = 0$ it is a “failure”
- $p_i = Pr(Y_i = 1)$
- If there are more 1s than 0s, then p_i should be higher than 0.50
- Can use other information x_i to influence the value of p_i in the model

Motivating Example

Suppose we want to model what effects a juror's verdict:

- $Y_i = 1$ means "guilty" and $Y_i = 0$ means "not guilty"
- Let x_i be a variable indicating whether or not the defendant's DNA was found at the crime scene. x_i is also binary: $x_i = 1$ when the defendant's DNA was found at the crime scene, and $x_i = 0$ otherwise
- We want to tie the probability a juror's verdict is guilty (p_i) to the presence of DNA evidence
- The probability should go up when there is DNA evidence, and should go down when there isn't DNA evidence.

Formulae

$$Y_i \sim \text{Bernoulli}(p_i)$$
$$\text{logit}(p_i) = \alpha + \beta \cdot x_i$$

Y_i is a random variable with $P(Y_i = 1) = p_i$, and the value of p_i changes with the with the value of other information x_i. α is the intercept, β is the slope of the model (similar to simple linear regression).

Logit???

The logit function takes a number from (0,1) (here, p_i) and turns it into a real number $(-\infty, \infty)$ (here $\alpha + \beta \cdot x_i$).

The inverse of this function (take a number from $(-\infty, \infty)$ and turn it into a number from (0,1)) is:

$$p_i = \frac{\exp\{\alpha + \beta \cdot x_i\}}{1 + \exp\{\alpha + \beta \cdot x_i\}}$$

This is why we use it for logistic regression: we can turn any value and combination of predictor variables into a probability in this way.

Logit example

Email data:

spam	to_multiple	winner
0	0	0
0	0	0
0	0	0
0	1	0
0	1	0
1	0	1

Is it spam? We think that this can be predicted by: whether or not it is sent to multiple people and/or it contains the word "winner"

Logit example

```r
glm(spam ~ to_multiple + winner, family = binomial, data = email)
```

```
##
## Call:  glm(formula = spam ~ to_multiple + winner, family = binomial,
##     data = email)
##
## Coefficients:
## (Intercept)  to_multiple       winner
##      -2.160       -1.802        1.502
##
## Degrees of Freedom: 3920 Total (i.e. Null);  3918 Residual
## Null Deviance:       2437
## Residual Deviance: 2349  AIC: 2355
```

Logit example

Table 2: Predicted probabilities of spam given the to_multiple and winner variables

spam	to_multiple	winner	n	pred_prob
0	0	0	2909	0.103
0	0	1	37	0.341
0	1	0	601	0.019
0	1	1	7	0.079
1	0	0	335	0.103
1	0	1	20	0.341
1	1	0	12	0.019